Virus Evolution, 2016, 2(2): vew024

doi: 10.1093/ve/vew024
Research article

Mapping the genomic diversity of HCV subtypes 1a and 1b: Implications of structural and immunological constraints for vaccine and drug development

Lize Cuypers,[1,*,†,§] Guangdi Li,[1,2,§] Christoph Neumann-Haefelin,[3] Supinya Piampongsant,[1,4] Pieter Libin,[5,1] Kristel Van Laethem,[1] Anne-Mieke Vandamme,[1,6,‡] and Kristof Theys[1]

[1]KU Leuven, University of Leuven, Department of Microbiology and Immunology, Rega Institute for Medical Research, Clinical and Epidemiological Virology, Minderbroedersstraat 10, 3000 Leuven, Belgium, [2]Metabolic Syndrome Research Center, Key Laboratory of Diabetes Immunology, Ministry of Education, National Clinical Research Center for Metabolic Diseases, the Second Xiangya Hospital, Central South University, Changsha, Hunan, China, [3]Department of Medicine II, Freiburg University Medical Center, University of Freiburg, Freiburg, Germany, [4]Department of Electrical Engineering ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, University of Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium, [5]Artificial Intelligence Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium and [6]Center for Global Health and Tropical Medicine, Microbiology Unit, Institute for Hygiene and Tropical Medicine, University Nova de Lisboa, Rua da Junqueira 100, Lisbon, 1349-008, Portugal

*Corresponding author: E-mail: lize.cuypers@kuleuven.be

†http://orcid.org/0000-0002-9433-8752

‡http://orcid.org/0000-0002-6594-2766

§These authors contributed equally to this work.

Abstract

Despite significant progress in hepatitis C (HCV) treatment, global viral eradication remains a challenge. An in-depth map of its genome diversity within the context of structural and immunological constraints could contribute to the design of pan-genotypic antivirals and preventive vaccines. For such analyses, extensive information is only available for the highly prevalent HCV genotypes (GT) 1a and 1b. Using 647 GT1a and 408 GT1b full-genome sequences obtained from the Los Alamos database, we found that respectively 3 per cent and 82 per cent of all codon positions are under positive and negative selective pressure, suggesting variation mainly accumulates due to random genetic drift. An association between conservation and both structured RNA and secondary protein structures confirmed the important role of structural elements at nucleotide and at amino acid level. Remarkably, $CD8^+$ T-cell epitopes in HCV GT1a were significantly more conserved, while at the same time containing more sites under positive selection. Similarly, $CD4^+$ T-cell epitopes were significantly more conserved in both HCV subtypes, but under less positive selective pressure in GT1b and more negative selective pressure in GT1a. In contrast, B-cell epitopes in both subtypes were less conserved and under less stringent negative selection. These findings argue against immune selective pressure as the main force of between-host diversifying evolution. Despite its high

variability, HCV is under strict evolutionary constraints, most probably to keep its genes and proteins functional during the replication cycle. These are encouraging findings for vaccine and drug design, which could consider these newly established genetic diversity profiles.

Key words: HCV GT1a and GT1b; conservation; selective pressure; RNA structure; T- and B-cell epitopes; protein secondary structure

1. Background

Hepatitis C virus (HCV) infections are associated with life-threatening diseases such as liver cirrhosis and hepatocellular carcinoma, remaining a global threat to public health (Webster et al. 2015). This enveloped RNA virus is classified into seven HCV genotypes (GTs) and more than 50 subtypes (Smith et al. 2014). HCV's single-stranded RNA genome encodes a large polyprotein with a single open-reading frame that translates into four structural and six non-structural (NS) proteins (Weiser and Tellinghuisen 2102).

To date, global eradication of HCV infections is challenged by the lack of a preventive vaccine that induces broad reactive immunity (Liang 2013). Two main approaches have been proposed in vaccine development, with the first targeting the humoral immune response as neutralizing antibodies (NAbs) play a major role in the clearance of HCV infections (Drummer 2014). However, a recombinant E1/E2 candidate tested in humans resulted into relatively low titres of NAbs, modest levels of neutralization and limited cross-neutralizing potential (Ray et al. 2010). Studies in humans and chimpanzees showed that T-cell responses are multifunctional and sustained over time, evolving towards the generation of T-cell-based vaccines (Liang 2013). The first T-cell-based vaccine with NS3-NS5B genes incorporated resulted into high levels of both $CD8^+$ and $CD4^+$ HCV-specific T cells targeting multiple HCV antigens, in humans (Swadling et al. 2014), yet no reactive immunity was established towards all circulating genotypes and subtypes.

Treatment progress is much more encouraging. Current anti-HCV treatment strategies rely on the administration of direct-acting antivirals (DAAs) that target HCV NS3/4A protease, NS5A or NS5B polymerase, and offer virological response rates of 95 per cent and higher for all circulating HCV genotypes (Yau and Yoshida 2014). Nevertheless, there is still a need to develop potent pan-genotypic drugs, characterized by minimal drug–drug interactions, affordable price tags and activity against HCV variants associated with drug resistance to current DAAs (Franco et al. 2014). Even with such high treatment successes, global eradication is not to be expected soon. To achieve viral eradication, the number of undiagnosed HCV patients urgently needs to be reduced (Mathis 2012; Wei and Lok 2014), which could enhance treatment coverage and prevent further transmission. High-risk and high-prevalence patient groups need to be screened with priority, given their high reported re-infection rates (Micallef et al. 2007; Martin et al. 2013; De Vos and Kretzschmar 2014).

The needed progress in vaccine design and therapeutic cure is hampered by the high evolutionary rate of HCV, which allows rapid adaptation to changing environments (Forns et al. 1999; Bailey et al. 2012; Li et al. 2015). Different molecular mechanisms have been proposed to shape viral evolution in general, including natural selection by the host immune system, selective pressures exerted by antiviral treatment, replication capacity, recombination, epidemiological factors such as migration, genetic bottlenecks and genetic drift, for which the two latter are largely influenced by population size (Preciado et al. 2014). Focusing on HCV, molecular evolution involves multiple complex processes characterized by temporal variations that constantly rearrange the architecture of the intra-host viral population. More specifically, viral variants can escape from strong selective pressure exerted by the host innate and adaptive immune system (Neumann-Haefelin and Thimme 2011; Heim and Thimme 2014). Under drug selective pressure, the heterogeneous intra-patient viral population can trigger the development of drug-resistant strains, with the resulting changes not equally distributed across the genome (Martell et al. 1992; Le Guillou-Guillemette et al. 2007; Gray et al. 2011). The replication capacity in a specific environment of a virus carrying a specific amino acid variant will influence whether the variant gets lost or becomes fixed in the viral population (Vandamme 2009; Neumann-Haefelin and Thimme 2011; Rehermann 2013), leading to genetic conservation in certain viral regions and adaptive evolution in others (Forns et al. 1999; Holmes 2003).

A previous in-depth analysis of genetic diversity and selective pressure in HCV genotypes 1–6 has shown that 39 per cent of all positions in the HCV full-genome had a consensus residue with a frequency of at least 95 per cent common for all genotypes (Cuypers et al. 2015). Positive selective pressure was identified for only 0.23 per cent of the HCV full-genome positions (Cuypers et al. 2015). The current study aims to advance these findings by integrating information on nucleotide and amino acid conservation, positive and negative selective pressure, and structural and immunological constraints. More specifically, a comprehensive map of structured RNA, protein secondary structures and immune epitopes will help to better understand viral evolution and how diversity is shaped by the host immune system and the tolerance to mutations (Gray et al. 2011; Snoeck et al. 2011; Bittar et al. 2013). These analyses require large datasets, hence our focus on the two most prevalent subtypes of the widely distributed HCV genotype 1.

2. Methods

2.1 Full-length genome sequence dataset

Full-length sequences (9,000 nt or more) of HCV GT1a and GT1b were downloaded from the Los Alamos HCV Sequence Database (LANL, http://hcv.lanl.gov) (Kuiken et al. 2008). Duplicates and sequences from non-human hosts were deleted, with a single sequence randomly selected per patient. Since all sequences of this study were submitted to Los Alamos before 2008, patients were considered to be DAA-naive, despite limited availability of treatment information.

Sequences were aligned using reference sequence H77, with an *in-house* developed pairwise alignment tool-chain (Cuypers et al. 2015; Libin 2014). Alignments were manually edited in Seaview V4.0 (Gouy et al. 2010) to improve their quality. HCV subtype assignment was checked based on clustering with

reference sequences from Los Alamos using a maximum-likelihood (ML) tree constructed by RAxML V8.0.20 (Stamatakis 2014). Four sequences showed a discordant result compared to their genotype assignments in LANL, confirmed by the COMET (Struck et al. 2014) and Oxford HCV subtyping tools (Alcantara et al. 2009), resulting into removal or inclusion of the sequence according to the genotype assigned by the ML tree and subtyping tools. Sequences containing in-frame stop codons or lacking sequence information for any of the HCV proteins were removed, resulting into a dataset of 647 HCV GT1a and 408 HCV GT1b full-genome sequences.

2.2 Representativeness of the dataset

The representativeness of the GT1a and GT1b full-genome datasets was evaluated with respect to virus diversity of the worldwide HCV GT1a and 1b epidemic. For each protein, probability distributions of pairwise nucleotide diversity calculated using HCV GT1a and GT1b full-genome datasets were compared with those of an evaluation dataset downloaded from the LANL database (Mann–Whitney *U* test, $P < 0.05$). This evaluation dataset contained all HCV GT1a and GT1b sequences present in Los Alamos, irrespective of any specific protein or genomic region, with only one sequence randomly selected per patient. Additionally, the full-genome datasets and evaluation datasets were compared by phylogenetic analysis using a ML approach.

2.3 Conserved residues and sites under positive or negative selective pressure

Within each subtype, the most prevalent amino acid or nucleotide at each position was defined as the consensus residue, even when its frequency was less than 50 per cent (Roebuck 2011). A position was defined as weakly conserved, conserved or highly conserved, when the frequency (x) of a consensus residue was <95 per cent, 95 per cent $\leq x <$ 99 per cent, or ≥99 per cent, respectively (Supplementary Table S1). Detection of selective pressure was performed using the fixed-effect likelihood (FEL) method, implemented in HyPhy v2.2.1 (Pond et al. 2005), which accommodates site-by-site variation. An amino acid position was considered to be under positive selective pressure if the ratio of non-synonymous and synonymous amino acid substitutions at that codon (dN/dS ratio) was >1 and the $P < 0.05$ as significance level for the likelihood ratio test (Supplementary Table S2). Contrary, amino acid positions with dN/dS <1, and a $P < 0.05$, were defined as positions under negative selective pressure.

2.4 Structural and immunological constraints

Linear HCV epitopes of T and B cells from assays that were able to identify epitopes were collected according to the Immune Epitope Database (IEDB) (http://www.iedb.org/) and mapped to their corresponding positions based on the reference sequence H77. T-cell epitopes restricted to the major histocompatibility complex (MHC) class I and II were analysed separately since $CD4^+$ T cells may not drive viral escape mutations (Fuller et al. 2010). HCV genotype 1 epitopes were only included when they were reported in humans. Moreover, only $CD8^+$ and $CD4^+$ T-cell epitopes described by at least two independent research groups were considered. Nevertheless, all B-cell epitopes were included because very few epitopes were reported by more than one research lab. Subtype information was not provided for many subjects in these studies, resulting in potential misclassification. Epitope annotations without subtype specification were allocated to both HCV subtype datasets.

A consensus RNA structure was predicted for each HCV subtype by the free energy minimization approach implemented in RNAalifold (Bernhart et al. 2008), based on ten random datasets of the GT1a and GT1b sequence alignments (mean, SD) because the maximum input of this software was set at 300 sequences. For each nucleotide position, RNAalifold assigns either a 'stem' or a 'loop', with nucleotide pairings within stems clearly indicated. Regions with more than five consecutive residues predicted as a stem were considered as a structured RNA region, without considering other secondary RNA structures (Jossinet et al. 2007; Kuznetsov et al. 2008).

Secondary structures for each of the ten HCV proteins were predicted based on either the sequence-based prediction software PSIPRED (Jones 1999; Buchan et al. 2013) or the structure-based prediction software 2Struc (Klose et al. 2010), at least when crystallized or nuclear magnetic resonance (NMR) structures were not present in the Protein Data Bank (PDB) (www.rcsb.org; Herman et al. 2000). Supplementary Table S3 summarizes the coverage of proteins by PDB data and their sequence similarity in comparison to reference sequence H77 (NC_004102). Protein secondary structures of regions not covered by PDB data were predicted with 2Struc or PSIPRED.

2.5 Generalized linear regression analysis

Given the datasets of the HCV GT1a and GT1b full-genomes and individual viral proteins, general linear regression analysis was performed to identify associations (1) between nucleotide conservation and structured RNA, (2) between amino acid conservation and the presence of epitopes or protein secondary structures, (3) between positive selected positions and immunological or structural constraints, (4) between negative selective pressure and immunological and structural constraints, and (5) between amino acid conservation and selective pressure. Odds ratios were estimated with a 95 per cent confidence interval, with an odds ratio considered as significant if its *P* was <0.05 after multiple testing corrections using the Benjamin–Hochberg method (Benjamini and Hochberg 1995). For multivariate analysis, sequence conservation or positive/negative selective pressure was considered as dependent variables and the different constraints as independent variables.

2.6 Resistance

Natural variation at positions associated with resistance development to drugs of the three DAA drug classes was studied more in depth in the context of structural and immunological constraints. Based on *in vitro* and/or *in vivo* data, drug resistance amino acid positions have been identified for all NS3/4A protease inhibitors ($n = 15$), NS5A inhibitors ($n = 9$), and NS5B polymerase inhibitors ($n = 14$) (Cuypers et al. 2015). The term resistance-associated variant (RAV) is used rather than the recently introduced term resistance-associated substitution, to indicate that natural variation at population level is studied, and no assumptions are made on the actual substitution. However, all variants are identified with respect to the generally accepted reference strains (H77 for GT1a and Con1 for GT1b). As such, Q80K means, for example, the 80K variant for which the reference strain has the amino acid Q and does not imply a substitution from Q to K.

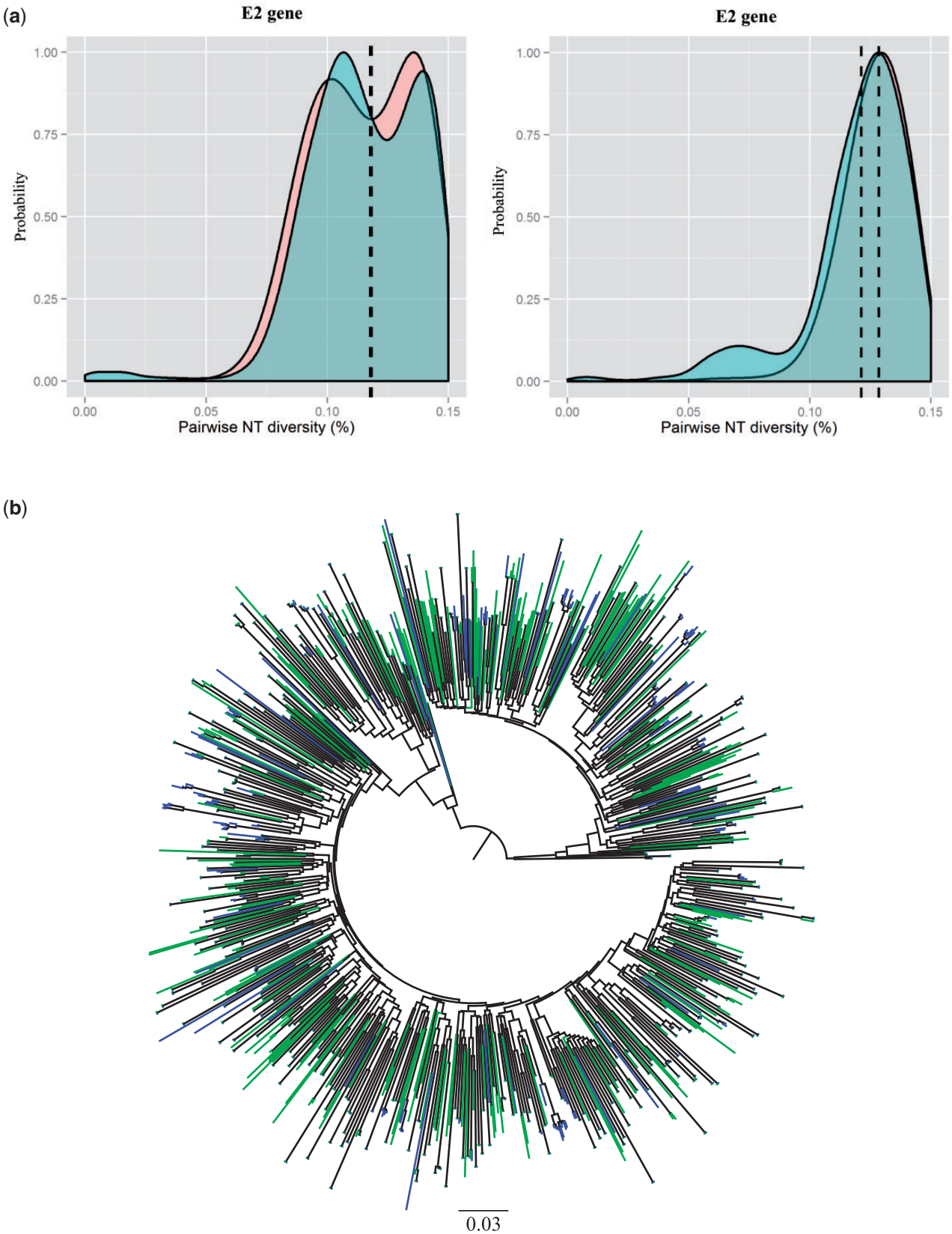


Figure 1. Representativeness of datasets.(A) The comparison of pairwise nucleotide (nt) diversity for the variable protein E2 of HCV GT1a (left) and HCV subtype GT1b (right). A strong overlap is shown between the full-genome dataset used (red), and the validation dataset consisting of all available sequences coding for gene E2 (blue). Similar overlaps in density plots were obtained for the different genome regions (not shown here). (B) HCV GT1a E2 sequences from the study (green), and validation (blue) datasets were aligned together and analysed in a bootstrapped (n = 1,000) ML tree using RaxML. Strains with an identical identification are displayed in black. The genetic distance is indicated underneath the tree. Similar results were obtained for other viral regions, as well as for HCV GT1b.

3. Results

3.1 Representativeness of the dataset

Large overlaps in diversity distributions were observed between the full-genome datasets and an evaluation dataset combining all available sequences for all proteins (Fig. 1A). Additionally, similarities were identified in the dispersion between strains from the full-genome dataset and the strains of the evaluation dataset (Fig. 1B). This suggests that both HCV GT1a and GT1b full-genome datasets can be considered as representative for the HCV GT1a and 1b sequence diversity in the complete Los Alamos database.

3.2 Conserved residues and positively selected positions

Only 15.7 per cent of the amino acid positions in the full-genome were defined as weakly conserved (Supplementary Table S1). Conversely, 60.8 per cent and 23.5 per cent were defined as, respectively, conserved and highly conserved within HCV GT1a, in comparison to 15.9 per cent and 68.5 per cent within HCV GT1b. At the nucleotide level, a higher proportion of weakly conserved positions was observed (26.4% for HCV GT1a and 29.0% for GT1b).

The FEL method predicted positive selective pressure at, respectively, 3.3 per cent and 2.9 per cent of the full-genome positions in HCV GT1a and GT1b (Supplementary Table S2). According to the codon-specific approach in FEL, 82.8 per cent of all positions across the HCV GT1a and GT1b full-genome were detected to be under negative selective pressure.

3.3 Mapping structural and immunological constraints

Figures 2 and 3, respectively, illustrate for each HCV GT1a and GT1b position the extent of nucleotide and amino acid diversity, detection of positive selective pressure, and different structural and immunological constraints.

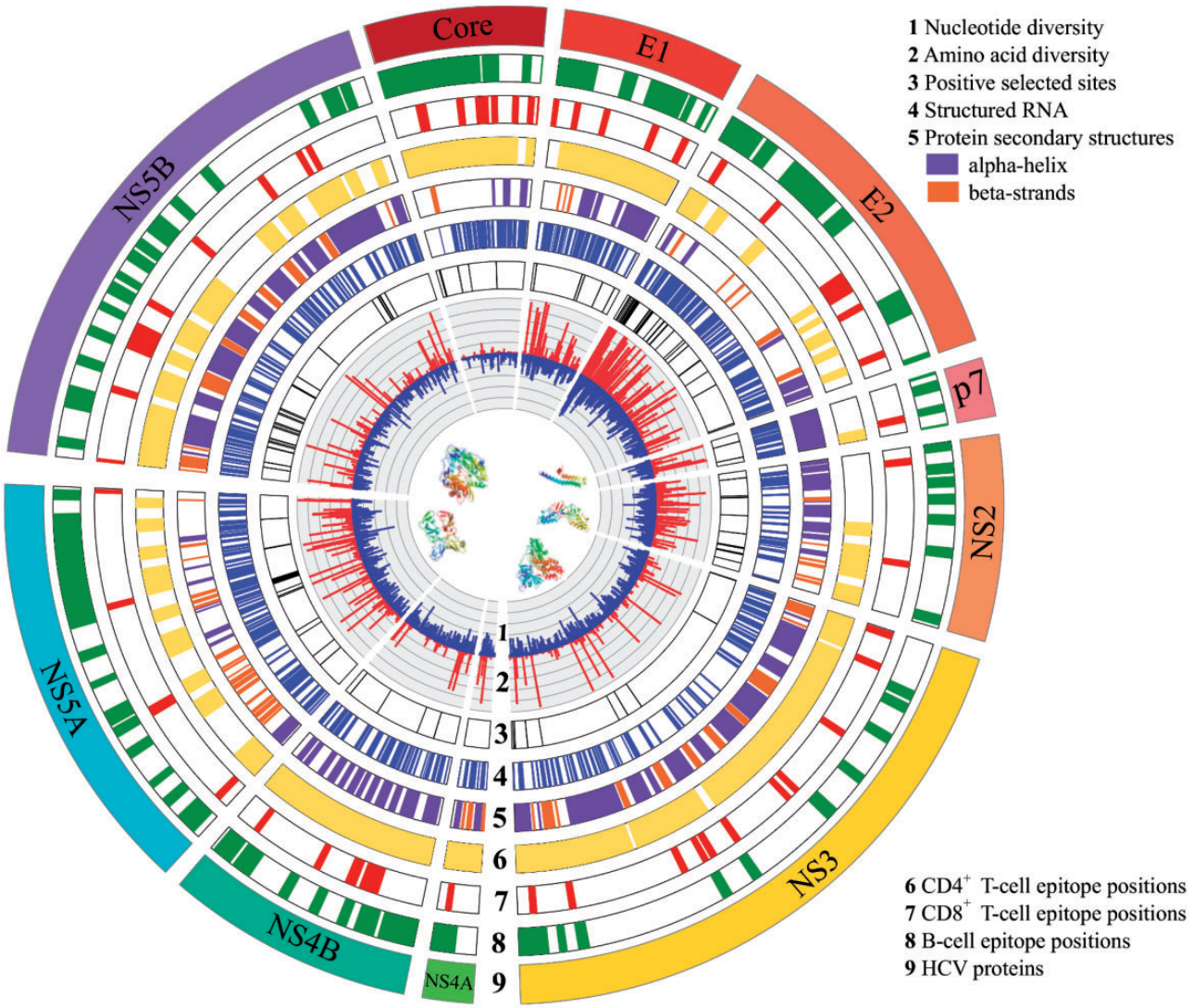


Figure 2. Mapping of sequence diversity, positive selective pressure, and structural and immunological constraints on the HCV GT1a full-genome. Layers: (1) average nucleotide diversity (per bar: 20%); (2) amino acid diversity (per bar: 20%); (3) positively selected sites; (4) RNA structures with five consecutive nucleotides predicted as "stem"; (5) protein secondary structures: α-helices (purple) and β-strands (orange); (6) $CD4^+$ T-cell epitope positions; (7) $CD8^+$ T-cell epitope positions; (8) B-cell epitope positions; (9) HCV proteins (reference strain: H77, NC_004102). Inner circle: 3D structures of proteins for which PDB resolution covers more than 50% of the protein (PyMOL V1.5 (http://www.pymol.org/), Circos V0.62 (http://circos.ca/; Krzywinski et al. 2009).

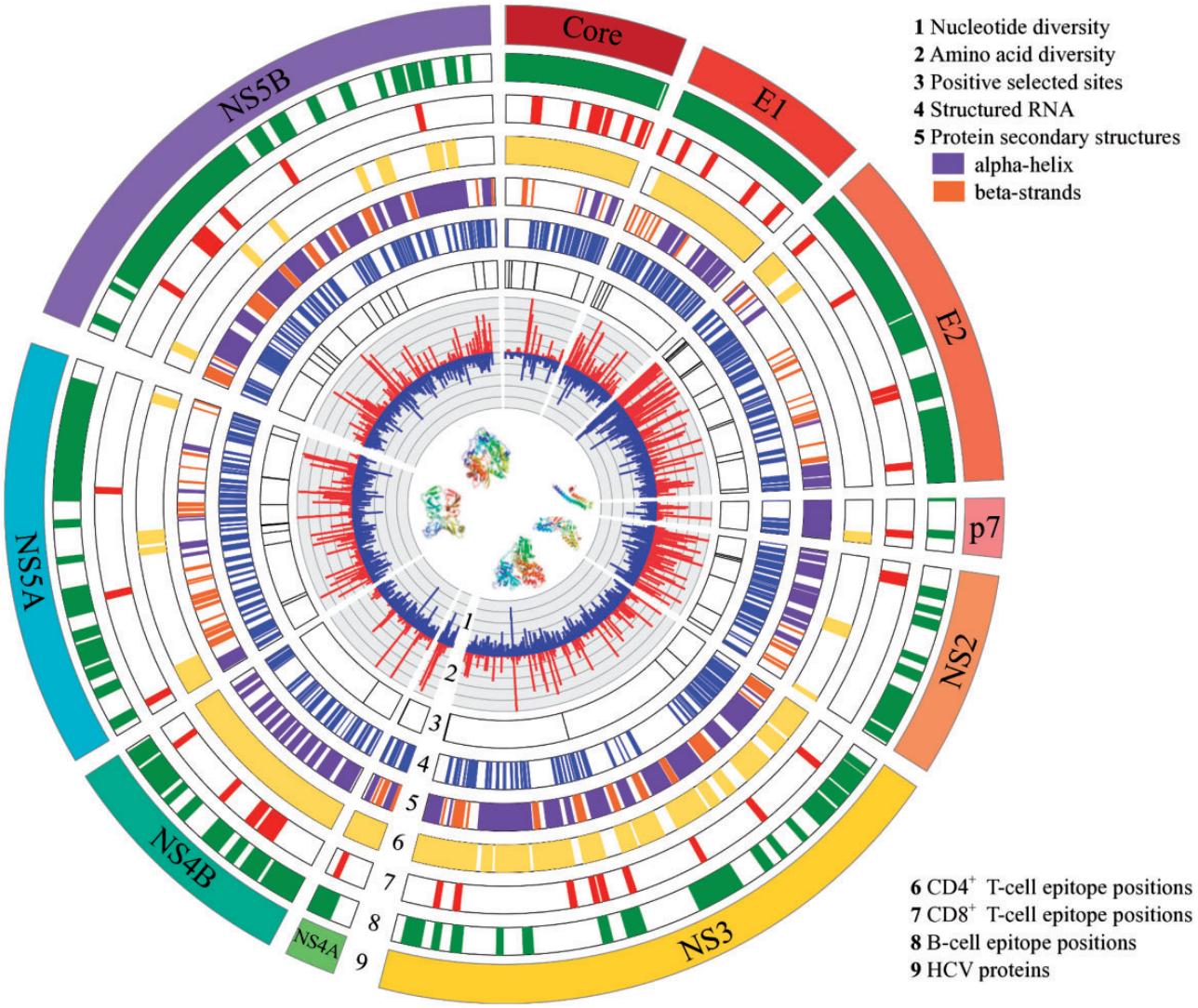


Figure 3. Mapping of sequence diversity, positive selective pressure, structural and immunological constraints, on the HCV GT1b full-genome. Same layers as in Figure 2.

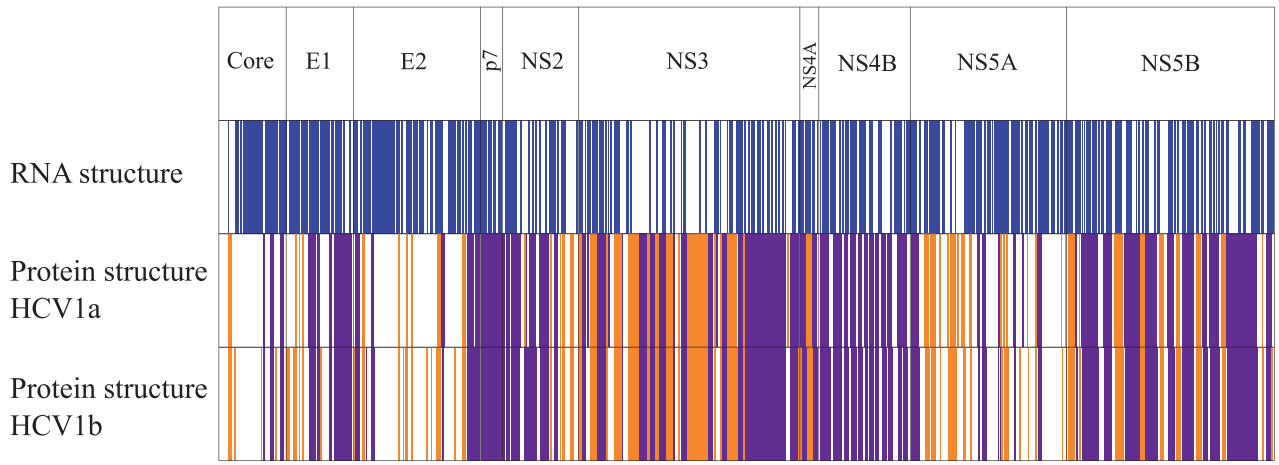


Figure 4. Alignment of structural constraints on the HCV GT1a vs HCV GT1b genome. The HCV genome is represented as a linear concatenation of all ten viral proteins (reference strain: H77), showing data for structured RNA regions (blue) and for protein secondary structures as α-helices (purple) or β-strands (orange).

Structured RNA regions, containing more than five consecutive residues predicted as a stem, were evenly distributed across the entire genome, covering 48 per cent and 46 per cent of the HCV GT1a and GT1b genome, respectively (Fig. 4). Of all viral proteins, protein E2 contained the lowest proportion of nucleotide positions in structured RNA regions (39.52% ± 0.39%), and protein NS4A the highest proportion (72.72% ± 3.33%), for both HCV subtypes.

The lowest proportion of protein secondary structures (α-helix and β-strand structures) was observed for the core protein (21.6% ± 2.55%), E2 (28.7% ± 3.89%), and NS5A (32.1% ± 0.64%) proteins (Fig. 4). On the contrary, more than 81 per cent of the amino acid positions of proteins NS5B and p7 were located within α-helix or β-strand structures. The distribution was very similar for both subtypes.

Overall, for $CD8^+$ T-cell epitopes, no significant difference in the number of mapped positions was observed between both HCV subtypes, in contrast to a higher number of $CD4^+$ T-cell epitopes in HCV GT1a and a higher number of B-cell epitopes for HCV GT1b. Within each subtype, B-cell epitopes and $CD4^+$ T-cell epitopes covered more positions than $CD8^+$ T-cell epitopes (49% and 73% vs. 17% for HCV GT1a, 67% and 50% vs. 17% for HCV GT1b). For both B-cell epitopes and $CD4^+$ T-cell epitopes, the highest proportion of positions was mapped in the core protein (Fig. 5). For $CD4^+$ T-cell epitopes, in addition to the core protein, high coverage of positions in proteins E1, NS3, NS4A, and NS4B was also observed. HCV proteins with the lowest coverage of epitopes were NS2 and NS5A for $CD8^+$ T-cell epitopes, E2, p7 and NS2 for $CD4^+$ T-cell epitopes, and p7 and NS3 for B-cell epitopes.

3.4 Combining different layers of data: univariate and multivariate analyses

For both HCV GT1a and GT1b, associations of sequence conservation or selective pressure with structural or immunological constraints were studied at the genome and individual protein level (Supplementary Tables S4A–C).

3.4.1 Univariate analysis: conservation

Supplementary Table S4A shows the associations between conservation and structural or immunological constraints. At the full-genome level, structured RNA regions displayed a significant positive correlation with nucleotide conservation in both subtypes (OR(GT1a): 1.13; OR(GT1b): 1.14). Moreover, this correlation was also significant for the individual core protein in the HCV GT1b dataset and for NS5B in the GT1a dataset. Across the HCV full-genome, B-cell epitopes and amino acid conservation were negatively correlated (OR(GT1a): 0.49; OR(GT1b): 0.59). This negative correlation was also significant for individual proteins: E2, NS5A, and NS5B for HCV GT1a and NS5A for GT1b strains. For both subtypes, a positive correlation between amino acid conservation and $CD4^+$ T-cell epitopes was observed (OR(GT1a): 1.82; OR(GT1b): 1.74) when using the entire genome, and this was confirmed for protein E2 in GT1b. In the full-genome of HCV GT1a, the presence of β-strands was positively correlated with amino acid conservation (OR: 1.37).

As expected, amino acid conservation displayed a significant negative correlation with positive selective pressure (OR(GT1a): 0.39; OR(GT1b): 0.38), as well as a positive correlation with negative selective pressure (OR(GT1a): 8.12; OR(GT1b): 4.79). Both correlations remained significant for the majority of the viral proteins separately (Supplementary Table S4A).

3.4.2 Univariate analysis: positive or negative selective pressure

Supplementary Table S4B shows associations between positive selection and different constraints. There was a significant low proportion of positions under positive selective pressure in structured RNA regions for the HCV GT1a full-genome (OR(GT1a): 0.65) and in α-helices for GT1b (OR(GT1b): 0.27). Of note, analyses on individual proteins yielded no significant results.

Supplementary Table S4C shows correlations between negative selective pressure and different constraints, which are more pronounced compared to correlations with positive selective pressure. In both HCV subtypes, β-strands had a significant enrichment of amino acid positions under negative selective pressure (OR(GT1a): 1.83; OR(GT1b): 1.53), while the reverse was true for B cell (OR(GT1a): 0.53; OR(GT1b): 0.45) and $CD8^+$ T-cell epitopes (OR(GT1a): 0.73; OR(GT1b): 0.72). Additionally, for the HCV GT1a dataset, α-helices had a low number of positions characterized by $dN/dS < 1$, while $CD4^+$ T-cell epitopes and structured RNA regions had more sites under negative selective pressure.

3.4.3 Multiple factors shaping HCV genotype 1 genomic diversity

As shown in Figure 6A and Supplementary Table S5, multivariate analysis suggested that in both subtypes, the only independent predictor for lower amino acid conservation was coverage

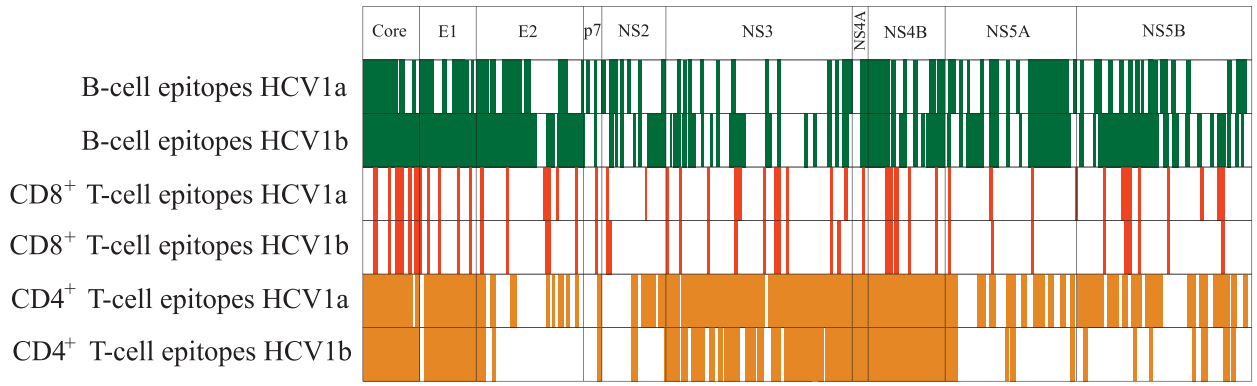


Figure 5. Alignment of immunological constraints on the HCV GT1a vs HCV GT1b genome. The HCV genome is represented as linear concatenation of the ten proteins (reference strain: H77), representing data on B-cell epitopes (dark green), $CD8^+$ T-cell epitopes (red) and $CD4^+$ T-cell epitopes (orange), for HCV GT1a and 1b separately.

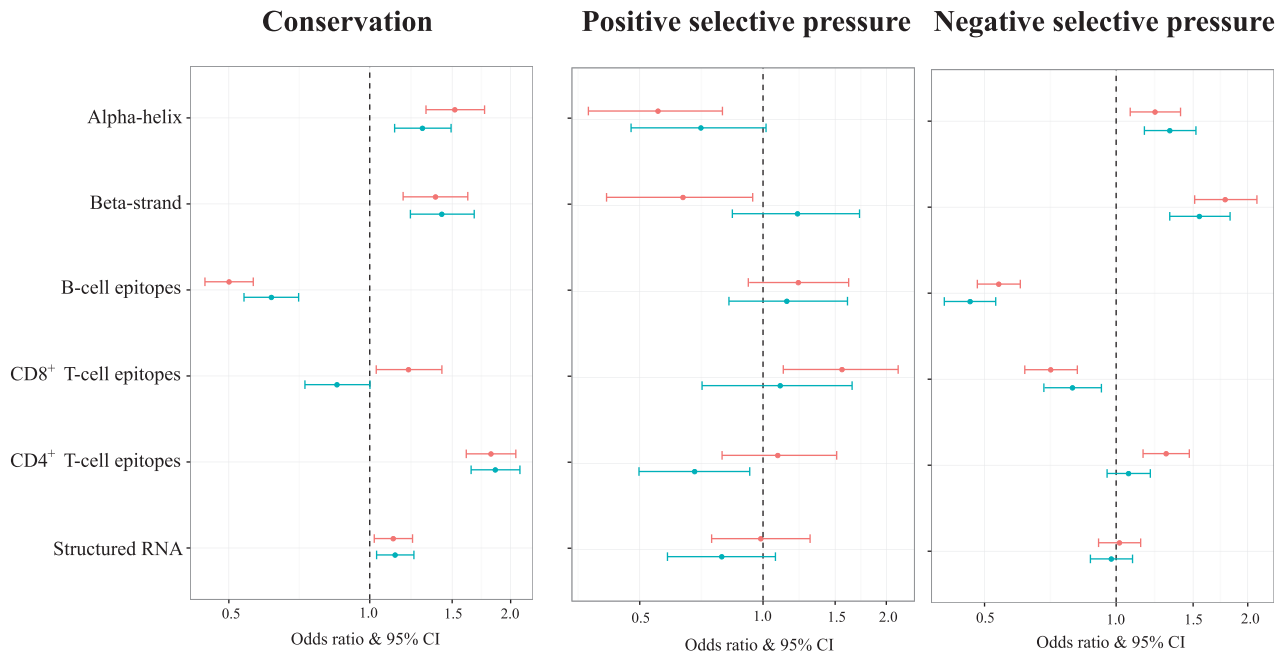


Figure 6. Multivariate analysis of residue conservation or selective pressure, with structural and immunological constraints. Multiple linear regression analysis was performed for nucleotide (structured RNA) or amino acid (all other comparisons) conservation (A, left), positive selective pressure (B, middle) and negative selective pressure (C, right). Odds ratios with 95 per cent confidence intervals (CI) are indicated for HCV GT1a (red) and HCV GT1b (blue).

by B-cell epitopes. There were, however, multiple independent predictors for higher conservation: coverage by $CD8^+$ T-cell epitopes, at least for HCV GT1a; $CD4^+$ T-cell epitopes; protein secondary structures (both α-helices and β-strands); and structured RNA regions. Predictors for a low number of sites under positive selective pressure were within α-helix and β-strand structures in the HCV GT1a full-genome (Fig. 6B). An independent predictor for a high number of sites under positive selective pressure was identified for HCV GT1a, $CD8^+$ T-cell epitopes, in contrast to $CD4^+$ T-cell epitopes for which the inverse relation was true for the GT1b dataset. A higher proportion of sites under negative selective pressure was identified in regions covered by α-helices and β-strands in both subtypes (Fig. 6C), while a lower proportion of sites was detected for B-cell epitopes and $CD8^+$ T-cell epitopes, also for both subtypes. Additionally, for the HCV GT1a full-genome, the presence of $CD4^+$ T-cell epitopes was an independent predictor for higher coverage by positions under negative selective pressure.

3.5 Resistance

For each of the drug resistance amino acid positions, structured RNA, $CD8^+$ and $CD4^+$ T-cell epitopes and B-cell epitopes, and protein secondary structures were mapped (Table 1). The majority of the fifteen drug resistance-related positions in protein NS3 was covered by structured RNA (10/15) or located within β-strand structures (12/15), expect one position embedded in an α-helix. More positions were targeted by B-cell compared to $CD8^+$ T-cell epitopes (Table 1A). Also for drug resistance-related positions in NS5A, a high number was covered by B-cell epitopes and mapped to structured RNA (Table 1B) and NS5B (Table 1C). In fact, this was a general phenomenon across all drug-targeted regions.

Additionally, the frequency distribution of amino acids at each amino acid position involved in drug resistance is indicated in Tables 1A–C. For NS3/4A protease inhibitors, a high natural prevalence was demonstrated for drug resistance-related variant Q80K in HCV GT1a-infected patients (39.3%). For NS5A inhibitors, several resistance-related variants had a relatively high natural prevalence, such as M28V for HCV GT1a (4.2%) and Y93H for HCV GT1b (3.9%). Moreover, NS5B variants C316N and S556G were present in 31.5 per cent and 8.4 per cent of the DAA-naive HCV GT1b-infected individuals.

4. Discussion

HCV's high evolutionary rate allows rapid adaptation to changing environments and subsequent escape from strong immune and treatment pressures, hampering both the development of vaccines and antiviral drugs. To investigate multiple driving factors of HCV genetic variability, this study provides a comprehensive map of HCV GT1a and GT1b full-genome genetic diversity integrating information on conservation, immune selective pressure, and immunological and structural constraints. The HCV subtypes 1a and 1b are the most prevalent worldwide with extensive availability of public data, giving our analysis sufficient statistical power. The full-genome datasets were found to be representative for the viral diversity observed worldwide in HCV GT1a and GT1b epidemics. Our findings contribute to the understanding of HCV evolution (Gray et al. 2011; Snoeck et al. 2011) and the interplay between HCV genetic diversity and host immunity (Bittar et al. 2013; Dolan et al. 2013).

A high proportion of genome positions in both subtypes was identified as conserved or highly conserved. Among all full-genome positions, only 3 per cent was found to be under positive selective pressure, predominantly found in the envelope glycoproteins. For both HCV GT1a and GT1b genomes, more than 82 per cent of all positions were observed to be under negative selective pressure, suggesting that variation mainly accumulates due to random genetic drift. At population level, HCV seems to follow the neutral model of evolution.

Table 1. Positions involved in drug resistance for NS3/4A protease inhibitors, NS5A inhibitors and NS5B polymerase inhibitors.

(A) NS3/4A protease inhibitor resistance-related positions

RAVs (H77)	V36MLA	F43SV	T54SA	V55A	Y56H	Q80KR	V107I	S122GR	I132V	R155K	A156STV	V158I	D168AVEYT	IV170AV	M175L
Natural prevalence HCV GT1a	$V^{97.4}$ $L^{1.2}$ $S^{1.1}$ **$M^{0.3}$**	$F^{98.8}$ $L^{1.1}$ $S^{0.1}$	$T^{97.1}$ $S^{1.7}$ $V^{1.1}$ $X^{0.1}$	$V^{96.1}$ $I^{2.0}$ $Y^{1.1}$ **$A^{0.8}$**	$Y^{98.8}$ $H^{1.1}$ $F^{0.1}$	$Q^{57.4}$ $K^{39.3}$ $D^{1.1}$ $N^{0.9}$ $L^{0.8}$ **$R^{0.3}$** $M^{0.2}$	$V^{98.9}$ $T^{1.1}$	$S^{94.1}$ $G^{4.6}$ **$R^{1.1}$** $X^{0.2}$	$I^{98.3}$ $S^{1.1}$ **$V^{0.5}$** $L^{0.1}$	$R^{98.3}$ $A^{1.1}$ **$K^{0.6}$**	A^{100}	$V^{98.9}$ $C^{1.1}$	$D^{98.6}$ $F^{1.1}$ **$E^{0.2}$** $G^{0.1}$	$I^{96.1}$ $V^{2.8}$ $P^{1.1}$	$L^{98.8}$ $E^{1.1}$ $X^{0.1}$
Natural prevalence HCV GT1b	$V^{99.0}$ **$L^{0.5}$** **$M^{0.3}$** **$I^{0.2}$**	F^{100}	$T^{97.8}$ **$S^{2.2}$**	V^{100}	$Y^{82.3}$ $F^{17.4}$ $X^{0.3}$	$Q^{95.6}$ $L^{3.7}$ $D^{0.3}$ **$K^{0.2}$** **$R^{0.2}$**	V^{100}	$S^{79.6}$ **$G^{12.3}$** **$T^{4.7}$** **$N^{3.4}$**	$V^{73.7}$ $I^{25.6}$ $L^{0.7}$	$R^{99.3}$ $P^{0.5}$ $X^{0.2}$	A^{100}	V^{100}	$D^{99.0}$ **$E^{1.0}$**	$V^{65.6}$ $I^{34.4}$	$M^{99.0}$ **$L^{1.0}$**

(B) NS5A inhibitors resistance-related positions

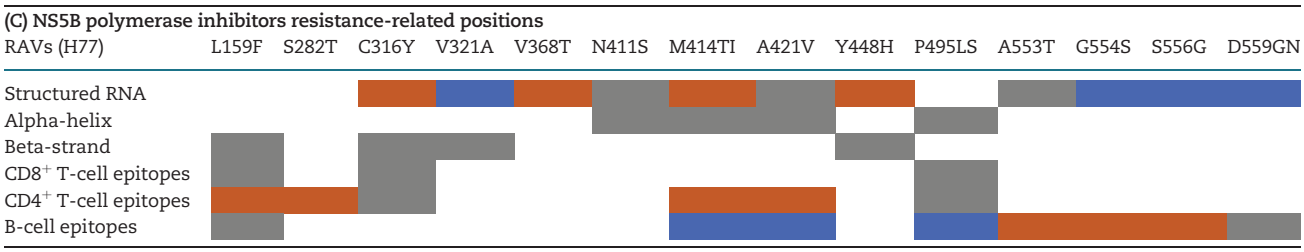
RAVs (H77)	M28TV	P29SX	Q30EHR	L31MV	P32LX	H58D	E62D	A92K	Y93HNC
Natural prevalence HCV GT1a	$M^{94.0}$ **$V^{4.2}$** $P^{1.1}$ **$T^{0.5}$** $I^{0.2}$	$P^{98.9}$ $Q^{1.1}$	$Q^{97.5}$ $L^{1.2}$ **$H^{0.9}$** **$R^{0.4}$**	$L^{97.7}$ **$M^{1.1}$** $P^{1.1}$ $X^{0.1}$	$P^{98.9}$ $G^{1.1}$	$H^{94.3}$ $P^{2.9}$ $C^{1.7}$ **$D^{0.3}$** $N^{0.3}$ $Y^{0.3}$ $Q^{0.2}$	$E^{96.9}$ $D^{1.7}$ $I^{1.1}$ $A^{0.2}$ $X^{0.1}$	$A^{98.4}$ $Y^{1.1}$ $P^{0.5}$	$Y^{97.8}$ $T^{1.1}$ **$C^{0.6}$** **$H^{0.5}$**
Natural prevalence subtype GT1b	$L^{97.1}$ **$M^{2.5}$** $V^{0.2}$ $F^{0.2}$	$P^{99.8}$ **$X^{0.2}$**	$R^{91.4}$ **$Q^{7.1}$** $K^{1.0}$ $L^{0.3}$ $M^{0.2}$	$L^{96.3}$ **$M^{3.4}$** **$I^{0.3}$**	P^{100}	$P^{93.4}$ **$S^{3.4}$** $T^{1.7}$ $L^{0.7}$ $Q^{0.5}$ $A^{0.3}$	$Q^{96.1}$ $E^{2.0}$ $K^{0.7}$ $R^{0.5}$ $S^{0.3}$ $H^{0.2}$ **$D^{0.2}$**	$A^{97.3}$ $T^{1.2}$ $V^{0.7}$ $G^{0.5}$ $S^{0.3}$	$Y^{95.8}$ **$H^{3.9}$** $F^{0.3}$

(C) NS5B polymerase inhibitors resistance-related positions

RAVs (H77)	L159F	S282T	C316Y	V321A	V368T	N411S	M414TI	A421V	Y448H	P495LS	A553T	G554S	S556G	D559GN
Natural prevalence HCV GT1a	$L^{98.9}$ $I^{1.1}$	$S^{98.8}$ $G^{1.1}$ $R^{0.1}$	$C^{98.8}$ $G^{1.1}$ $X^{0.1}$	$V^{99.9}$ $I^{0.1}$	$S^{98.6}$ $N^{1.1}$ $\_^{0.3}$	$N^{98.6}$ $I^{1.1}$ $\_^{0.3}$	$M^{98.5}$ $F^{1.1}$ $\_^{0.3}$ $V^{0.1}$	$A^{84.5}$ **$V^{12.7}$** $\_^{1.2}$ $R^{1.1}$ $T^{0.3}$ $M^{0.2}$	$Y^{98.6}$ $G^{1.1}$ $\_^{0.3}$	$P^{98.4}$ $\_^{1.6}$	$A^{89.6}$ $\_^{9.8}$ $G^{0.6}$	$G^{89.8}$ $\_^{9.8}$ $Y^{0.4}$	S^{89} $\_^{9.8}$ **$G^{1.1}$** **$N^{0.1}$**	$D^{89.8}$ $\_^{9.6}$ $I^{0.6}$
Natural prevalence HCV GT1b	$L^{94.8}$ **$F^{5.2}$**	$S^{99.8}$ **$T^{0.2}$**	$C^{67.8}$ **$N^{31.5}$** $H^{0.3}$ $R^{0.2}$ **$Y^{0.2}$**	$V^{99.5}$ $I^{0.5}$	$S^{99.8}$ $P^{0.2}$	N^{100}	$M^{99.8}$ $I^{0.2}$	$A^{94.1}$ $V^{5.7}$ $S^{0.2}$	$Y^{99.5}$ $C^{0.3}$ **$H^{0.2}$**	P^{100}	$A^{99.5}$ $\_^{0.3}$ $V^{0.2}$	$G^{99.5}$ $\_^{0.3}$ $X^{0.2}$	$S^{88.9}$ **$G^{8.4}$** $N^{1.5}$ $D^{0.7}$ $\_^{0.3}$ $X^{0.2}$	$D^{98.3}$ $\_^{1.5}$ $N^{0.2}$

(continued)

Table 1. Continued



Respectively, 15, 9, and 14 positions were identified as being associated with drug resistance to (A) NS3/4A protease inhibitors, (B) NS5A inhibitors, and (C) NS5B polymerase inhibitors. For each of these positions, the constraints structured RNA, protein secondary structures and T- and B-cell epitopes were mapped. Positions covered by constraints in both datasets were colored in grey. When only mapped for HCV GT1a, the position was indicated in red and only for HCV GT1b sequences in blue. For each of the three drug classes, the natural prevalence of RAVs is listed, using H77 as reference. Frequencies are indicated in superscript, and the amino acid variants are ranked according to decreasing frequency. Amino acid substitutions known to confer drug resistance towards DAAs are indicated in bold, for both HCV subtypes.

4.1 Sequence conservation is higher in structured regions

Structured RNA regions were associated with nucleotide conservation in both HCV GT1a and GT1b (Heim and Thimme 2014), revealing that genomic structures play important roles during the viral life cycle, as has been described for the RNA internal ribosome entry site in the 5′-untranslated region (Filbin and Kieft 2011; Berry et al. 2011). The same association was also significant in the HCV GT1a dataset for the NS5B gene, consistent with a *cis*-acting replication element in NS5B whose conserved RNA structure is essential for bringing the 5′ and 3′ ends of the genome (You et al. 2004; Tuplin et al. 2012). Structured RNA regions were also negatively correlated with evidence of positive selective pressure, while positions under negative selective pressure were enriched for HCV GT1a, however not confirmed in the multivariate analysis. The need to maintain RNA structure restricts the possibilities for protein evolution, and amino acid changes as a result of positive selective pressure might disrupt RNA secondary structure (Tuplin et al. 2012). This shows that for RNA viruses, selective pressure may not be properly measured by only investigating amino acid changes. There is a need for methods to measure selective pressure at nucleotide level.

Multivariate analysis showed an enrichment of conserved amino acid residues in regions containing α-helix and β-strand structures. A negative association with positive selective pressure further illustrates the importance of both α-helices and β-strands, which was only found in HCV GT1a, while previously also reported for HIV (Canducci et al. 2009). Then again, negative selective pressure was significantly associated with both α-helices and β-strands in both subtypes. These findings are in agreement with both protein secondary structures being important for protein functions, in contrast to what was reported by Snoeck *et al.* for HIV (Snoeck et al. 2011). Maintaining secondary structure exerts a negative selective pressure on the genome to minimize amino acid changes that may impair protein functions (Shabalina et al. 2013). However, it is worth to mention that 3D structure information for HCV proteins is not so extended as for HIV proteins. Additionally, it needs to be noted that both for RNA and protein secondary structures, correlations with positively selected sites were only significant for HCV GT1a, potentially due to the larger amount of data available for this subtype.

4.2 B-cell and T-cell epitopes

The IEDB database contains more B-cell epitopes for HCV GT1b compared to GT1a, with only few epitopes reported by two or more research groups. The inclusion of all available B-cell epitopes could have overestimated the number of B-cell epitope positions in the HCV genome. Moreover, the majority of B-cell assays focuses on structural proteins, resulting into high coverage rates for the core protein and the E2 hypervariable region. For T-cell epitope investigations, a stricter selection was possible, MHC class I ($CD8^+$ T-cell epitopes)- and MHC class II ($CD4^+$ T-cell epitopes)-restricted epitopes were only taken into account when reported by two or more research groups. Most HCV-infected individuals display only few HCV-specific $CD8^+$ T-cell responses (Rehermann 2013), and thus, as expected, a lower number of positions is covered by $CD8^+$ T-cell epitopes than by B-cell epitopes in our analysis.

Both MHC class epitopes were analyzed separately since $CD4^+$ T cells are believed not to drive viral escape (Fuller et al. 2010), while $CD8^+$ T cells have been shown to contribute to viral escape (Bowen and Walker 2005). Until recently, a general HCV GT1a peptide set was used to analyze T-cell epitopes in individuals infected with any genotype, resulting into a higher quality of described epitopes in GT1a compared to GT1b. Unfortunately, not all studies in the database had HCV genotype information available, resulting in the allocation of HCV GT1 epitopes without subtype specification to both HCV subtype datasets, potentially introducing biases in the analysis.

For HCV proteins NS2 and NS5A, low coverage was observed for both T-cell epitopes, compared to B-cell epitopes. This difference might be explained by the fact that the NS2 protein is a 'short-lived' protein (Chevaliez and Pawlotsky 2006), and the discovery that protein NS5A impairs both the innate and adaptive immune response to promote chronic HCV infections (Kriegs et al. 2009). Moreover, the lowest numbers of B-cell epitope positions were observed for proteins p7 and NS3, while for NS3 a high coverage by $CD4^+$ T-cell epitopes was found. For p7, it is probably due to the unusual architecture of the ion channel, which harbours hydrophobic pockets (OuYang et al. 2013). For NS3, it might be due to the role of this protein in viral escape from adaptive immune responses (Horner and Gale 2013).

4.3 The host immune system may be not the driving force of HCV variability

Amino acid substitutions assist in viral escape from immunological responses, expecting B- and $CD8^+$ T-cell epitope positions to be under positive selective pressure (Holmes 2003). In the multivariate analysis, positively selected sites were enriched in $CD8^+$ T-cell epitopes of GT1a only. No significant

correlation was found for B-cell epitopes, while for $CD4^+$ T-cell epitopes, a negative correlation was observed, although only in GT1b. B- and $CD8^+$ T-cell epitopes were independent predictors of a lower proportion of positions under negative selective pressure, while for $CD4^+$ T-cell epitopes the correlation was the reverse, at least in GT1a. Together these results confirm the hypothesis that $CD4^+$ T cells are not driving viral escape (Fuller et al. 2010) and suggest that humoral immune selective pressure is unlikely to act as driving force of variation. Differently, $CD8^+$ T-cell immune response may have some impact on the HCV genome.

T-cell responses are much more likely to be identified in conserved viral regions since T-cell assays have been developed using conserved peptides, and this may explain in part why conserved residues were found to be enriched within $CD4^+$ and $CD8^+$ T-cell epitope positions. Nevertheless, the observation that conserved sites do contain epitopes is useful in the vaccine field as T-cell-based vaccines incorporating conserved segments are underway. For $CD4^+$ T-cell epitopes, a similar association has been observed for HIV (Sanjuán et al. 2013). The significant correlation of $CD8^+$ T-cell epitopes with both positive selective pressure and amino acid conservation argues that while immune cellular selective pressure does exist, it may not be the main force for diversifying evolution. For B-cell epitopes, a negative correlation was found with amino acid conservation.

Overall, these findings suggest that the host immune system may not be the driving factor of HCV genetic diversity, at least from a population perspective, given that we only analyzed between-host genetic diversity. A larger proportion of diversity due to immune selective pressure can be expected in the context of host immune genetic background (Kuniholm et al. 2010). This may be related to the fact that B cells that contain HCV receptors have been identified as natural reservoirs of HCV (Inokuchi et al. 2009; Ito et al. 2011), and HCV also infects T cells (Blackard et al. 2004; Kondo et al. 2007) although the exact mechanism is yet unknown.

4.4 Drug resistance-associated positions are found in highly structured regions

Drug RAVs in all three target genes, located near the active site of the proteins (Elfiky et l. 2013; Meeprasert et al. 2014; Nettles et al. 2014), were consistently found in regions with a high coverage of structured RNA regions and folding of protein β-strands. Additionally, a higher number of RAVs was mapped to B-cell compared to $CD8^+$ T-cell epitopes, which could be explained by higher abundance, yet active sites or enzymatic pockets of viral proteins would be expected to be less accessible for NAbs. All three drug targets harboured particular RAVs with a high natural prevalence, consistent with what has been reported before (Dietz et l. 2015; Lawitz et al. 2015; Costantino et al. 2015): NS3 variant Q80K (GT1a: 39.3%), NS5A variants M28V (GT1a: 4.2%), and Y93H (GT1b: 3.9%) and NS5B variants C316N (GT1b: 31.5%) and S556G (GT1b: 8.4%).

4.5 Limitations

The study only focused on HCV GT1a and GT1b given limitations in the number of viral sequences available for other HCV genotypes in the Los Alamos database and in the structural information present in the IEDB and PDB databases. For those HCV GT1a sequences for which sampling location was known, a large proportion was from the USA, potentially explaining the high prevalence of the NS3 natural variant Q80K in this study (McCloskey et al. 2015). Concerning data for B-cell epitopes, a bias was introduced since only linear epitopes were considered, and relaxed inclusion criteria were used compared to T-cell epitopes. Restricted information was available for 3D protein structures (Supplementary Table S3), necessitating structural predictions for regions lacking PDB data. Nucleotide positions assigned as stems were considered to be structured RNA regions, excluding other regions such as loops, junctions, helices, and bulges from the analysis.

5. Conclusions

This in-depth study of HCV GT1a and GT1b by mapping constraints onto the full-genome shows a strong sequence conservation of structured regions, even across sites mapped by immune epitopes, emphasizing the important role of genomic and protein structures in the viral life cycle. The selective pressure analysis indicates that the immune selective pressure present does not act as a main force for HCV between-host diversifying evolution, suggesting HCV diversity has been accumulating mainly due to random genetic drift. Despite its high genetic variability, HCV GT1 is under strict evolutionary constraints, most probably to preserve the inherent functions of HCV proteins such as their interactions with host proteins during the viral life cycle. Since the design of a broad reactive vaccine is challenged, the implementation of genotype-specific and even geographically tailored vaccine immunogens should be explored.

Acknowledgements

The authors wish to thank Paul Proost and Johan Van Weyenbergh for their participation in several valuable discussions. Special gratitude goes to Tim Dierckx for his assistance in the use of software package R for the visualization of the results.

Funding

Lize Cuypers and Pieter Libin were supported by a PhD grant of the FWO (Fonds Wetenschappelijk Onderzoek – Vlaanderen, Asp/12 and Asp/15); Guangdi Li by the National Natural Science Foundation of China (31571368) and by the project of Innovation-driven Plan of Central South University (2016CX031). Kristof Theys was sponsored by a postdoctoral grant of the FWO (PDO/11). Part of this research was sponsored by two FWO grants (G.A029.11N and 1S31916N) and a grant from the VUB (VUB/OZR2714). The computational resources and services used in this work were provided by the Hercules Foundation and the Flemish Government – department EWI-FWO Krediet aan Navorsers (Theys, KAN2012 1.5.249.12.). The authors declare that they have no other competing interests than the financial disclosures above.

Authors' Contributions

LC gathered the full-genome sequence data, performed all analyses and wrote subsequent drafts of the article. GL participated in the study design, the quality control of the data and the visualization of the results, and assisted in writing the article. SP co-operated with the performance of the analyses. PL provided software implementations and supported

the high performance computational needs of this project. CNH assisted in writing the article. Both KVL and AV participated in the study design and in writing the article. KT supervised the design of the study, provided support for the visualization of the results, and assisted in writing the article. All authors read and approved the final article.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Alcantara, L. C., et al. (2009) 'A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences', *Nucleic Acids Research*, 37: W634–42

Bailey, J. R., et al. (2012) 'Constraints on viral evolution during chronic hepatitis C virus infection arising from a common-source exposure', *Journal of Virology*, 86: 12582–90.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of Royal Statistical Society*, 57: 289–300

Bernhart, S. H., et al. (2008) 'RNAalifold: improved consensus structure prediction for RNA alignments', *BMC Bioinformatics*, 9: 474.

Berry, K. E., et al. (2011) 'Crystal structure of the HCV IRES central domain reveals strategy for start-codon positioning', *Structure*, 19: 1456–66.

Bittar, C., et al. (2013) 'On hepatitis C virus evolution: the interaction between virus and host towards treatment outcome', *PLoS One*, 8.

Blackard, J. T., et al. (2004) 'Hepatitis C virus (HCV) diversity in HIV-HCV-coinfected subjects initiating highly active antiretroviral therapy', *Journal of Infectious Disease*, 189: 1472–81.

Bowen, D. and Walker, C. (2005) 'Adaptive immune responses in acute and chronic hepatitis C virus infection', *Nature*, 436: 946–52.

Buchan, D. W., et al. (2013) 'Scalable web services for the PSIPRED Protein Analysis Workbench', *Nucleic Acids Research*, 41: W349–57.

Canducci, F., et al. (2009) 'Dynamic features of the selective pressure on the human immunodeficiency virus type 1 (HIV-1) gp120 $CD4^{+}$-binding site in a group of long term non progressor (LTNP) subjects', *Retrovirology*, 6: 4.

Chevaliez, S. and Pawlotsky, J. -M. (2006). 'HCV genome and life cycle'. In: S. L., Tan (ed.) *Hepatitis C Viruses: Genomes and Molecular Biology*. Norfolk (UK): Horizon Bioscience, Chapter 1.

Costantino, A., et al. (2015) 'Naturally occurring mutations associated with resistance to HCV NS5B polymerase and NS3 protease inhibitors in treatment-naïve patients with chronic hepatitis C', *Virol Journal*, 12: 186.

Cuypers, L., et al. (2015) 'In context of eradication of the hepatitis C virus: genetic diversity and selective pressures of HCV genotype 1-6', *Viruses*, 7: 5018–39.

De Vos, A. S. and Kretzschmar, M. E. E. (2014) 'Benefits of hepatitis C virus treatment: A balance of preventing onward transmission and re-infection', *Mathematical Biosciences*, 258: 8–11.

Dietz, J., et al. (2015) 'Consideration of viral resistance for optimization of direct antiviral therapy of hepatitis C virus genotype 1-infected patients', *PLoS One*, 10/8: e0134395.

Dolan, D. T., et al. (2013) 'Identification and comparative analysis of hepatitis C virus-host cell protein interactions', *Molecular Biosystems*, 9: 3199–209.

Drummer, H. E. (2014) 'Challenges to the development of vaccines to hepatitis C virus that elicit neutralizing antibodies', *Frontiers Microbiology*, 5: 329.

Elfiky, A. A., et al. (2013) 'Molecular modelling comparison of the performance of NS5B polymerase inhibitor (PSI-7977) on prevalent HCV genotypes', *Protein Jouranl*, 32: 75–80.

Filbin, M. E. and Kieft, J. S. (2011) 'HCV IRES domain IIb affects the configuration of coding RNA in the 40S subunit's decoding groove', *RNA*, 17: 1258–73.

Forns, X., Purcell, R. H., and Bukh, J. (1999) 'Quasispecies in viral persistence and pathogenesis of hepatitis C virus', *Trends in Microbiology*, 7: 402–10.

Franco, S., et al. (2014) 'Detection of a sexually transmitted hepatitis C virus protease inhibitor-resistance variant in a human immunodeficiency virus-infected homosexual man', *Gastroenterology*, 174: 599–601.

Fuller, M. J., et al. (2010) 'Selection-driven immune escape is not a significant factor in the failure of $CD4^{+}$ T cell responses in persistent hepatitis C virus infection', *Hepatology*, 51: 378–87.

Gouy, M., Guindon, S., and Gasuel, O. (2010) 'SeaView version 4: A multiplatform graphical user interace for sequence alignment and phylogenetic tree building', *Molecular Biology and Evolution*, 27: 221–4.

Gray, R. R., et al. (2011) 'The mode and tempo of hepatitis C virus evolution within and among hosts', *BMC Evolutionary Biology*, 11: 131.

Heim, M. H. and Thimme, R. (2014) 'Innate and adaptive immune responses in HCV infections', *Journal of Hepatology*, 61: S14–25.

Herman, H. M., et al. (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28: 235–42.

Holmes, E. C. (2003) 'Error thresholds and the constraints to RNA virus evolution', *Trends in Microbiology*, 11: 543–6.

Horner, S. M. and Gale, M. Jr. (2013) 'Regulation of hepatic innate immunity by hepatitis C virus', *Nature Medicine*, 19: 879–88.

Inokuchi, M., et al. (2009) 'Infection of B cells with hepatitis C virus for the development of lymphoproliferative disorders in patients with chronic hepatitis C', *Jorunal of Medical Virology*, 81: 619–27.

Ito, M., Kusunoki, H., and Mizuochi, T. (2011) 'Peripheral B cells as reservoirs for persistent HCV infection', *Frontiers inMicrobiology*, 2: 177.

Jones, D. T. (1999) 'Protein secondary structure prediction based on position-specific scoring matrices', *Journal of Molecular Biology*, 292: 195–202.

Jossinet, F., Ludwig, T. E., and Westhof, E. (2007) 'RNA structure: bioinformatic analysis', *Current Opinion in Microbiology*, 10: 279–85.

Klose, D. P., Wallace, B. A., and Janes, R. W. (2010) '2Struc: the secondary structure server', *Bioinformatics*, 26: 2624–5.

Krzywinski, M. I., et al. (2009) 'Circos: An information aesthetic for comparative genomics', *Genome Research*, 19: 1639–45.

Kondo, Y., et al. (2007) 'Hepatitis C virus infects T cells and affects interferon-gamma signaling in T cell lines', *Virology*, 361: 61–173.

Kriegs, M., et al. (2009) 'The hepatitis C virus non-structural NS5A protein impairs both the innate and adaptive hepatic immune response *in vivo*', *Journal of Biological Chemistry*, 284: 28343–51.

Kuiken, C., et al. (2008) 'The hepatitis C sequence database in Los Alamos', *Nucleic Acids Research*, 36: D512–6.

Kuniholm, M. H., et al. (2010) 'Specific HLA class I and II alleles associated with hepatitis C virus viremia', *Hepatology*, 51: 1514–22.

Kuznetsov, S. V., et al. (2008) 'Loop dependence of the stability and dynamics of nucleid acid hairpins', *Nucleic Acids Research*, 36: 1098–112.

Lawitz, E., et al. 2015. *A phase 3, open-label, single-arm study to evaluate the efficacy and safety of 12 weeks of simeprevir (SMV) plus sofosbuvir (SOF) in treatment-naïve or –experienced patients with chronic HCV genotype 1 infection and cirrhosis: OPTIMIST-2. 50th EASL*, Vienna, Austria, April 22-6, 2015. Abstract LP04.

Le Guillou-Guillemette, H., et al. (2007) 'Genetic diversity of the hepatitis C virus: impact and issues in the antiviral therapy', *World Journal of Gastroenterology*, 13: 2416–26.

Li, G., et al. (2015) 'An integrated map of HIV genome-wide variation from a population perspective', *Retrovirology*, 12: 18.

Liang, T. J. (2013) 'Current progress in development of hepatitis C virus vaccines', *Nature Medicine*, 19: 869–78.

Libin, P. 2014. Applying graphical modelling techniques to virological data. *Thesis for the fulfilment of Master of Science in Applied Informatics*, Vrije Universiteit Brussel, June 14, 2014.

Martin, T. C., et al. (2013) 'Hepatitis C virus reinfection incidence and treatment outcome among HIV-positive MSM', *Aids*, 27: 2551–7.

Mathis, A. S. (2012) 'Economic burden and current managed care challenges associated with hepatitis C', *American Journal of Managed Care*, 14/Suppl: S350–9

Martell, M., et al. (1992) 'Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution', *Journal of Virology*, 66: 3255–9

McCloskey, R. M., et al. (2015) 'Global origin and transmission of hepatitis C virus non-structural protein 3 Q80K polymorphism', *Journal of Infectious Disease*, 211: 1288–95.

Meeprasert, A., Hannongbua, S., and Rungrotmongkol, T. (2014) 'Key binding and susceptibility of NS3/4A serine protease inhibitors against hepatitis C virus', *Journal of Chemical Information and Modeling*, 54: 1208–17.

Micallef, J. M., et al. (2007) 'High incidence of hepatitis C virus reinfection within a cohort of injecting drug users', *Journal of Viral Hepatitis*, 14: 413–8.

Nettles, J. H., et al. (2014) 'Assymetric binding to NS5A by daclatasvir (BMS-790052) and analogues suggests two novel modes of HCV inhibition', *Journal of Medicinal Chemistry*, 57: 20031–10043.

Neumann-Haefelin, C. and Thimme, R. (2011) 'Success and failure of virus-specific T cell response in hepatitis C virus infection', *Digestive Diseases*, 29: 416–22.

OuYang, B., et al. (2013) 'Unusual architecture of the p7 channel from hepatitis C virus', *Nature*, 498: 521–5.

Pond, S. L., Frost, S. D., and Muse, S. V. (2005) 'HyPhy: hypothesis testing using phylogenies', *Bioinformatics*, 21: 676–9.

Preciado, M. V., et al. (2014) 'Hepatitis C virus molecular evolution: transmission, disease progression and antiviral therapy', *World Journal of Gastroenterology*, 20: 15992–6013.

Roebuck, K. (2011) 'Biochips: high-impact strategies – what you need to know: definitions, adoptions, impact, benefits, maturity, vendors', *Emereo Publishing*, 205–7 p

Ray, R., et al. (2010) 'Characterization of antibodies induced by vaccination with hepatitis C virus envelope glycoproteins', *Journal of Infectious Disease*, 202: 862–6.

Rehermann, B. (2013) 'Pathogenesis of chronic viral hepatitis: differential roles of T cells and NK cells', *Nature Medicine*, 19: 859–68.

Sanjuán, R., et al. (2013) 'Immune activation promotes evolutionary conservation of T-cell epitopes in HIV-1', *PLoS Biol*, 11.

Shabalina, S. A., Spiridonov, N. A., and Kashina, A. (2013) 'Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity', *Nucleic Acids Research*, 41: 2073–94.

Smith, D. B., et al. (2014) 'Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource', *Hepatology*, 59: 318–27.

Snoeck, J., et al. (2011) 'Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints', *Retrovirology*, 8: 87–94.

Stamatakis, A. (2014) 'RaxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30: 1312–3.

Struck, D., et al. (2014) 'COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification', *Nucleic Acids Research*, 42: e144.

Swadling, L., et al. (2014) 'A human vaccine strategy based on chimpanzee adenoviral and MVA vectors that primes, boosts, and sustains functional HCV-specific T cell memory', *Science Translational Medicine*, 6: 261ra153.

Tuplin, A., et al. (2012) 'A twist in the tail: SHAPE mapping of long-range interactions and structural rearrangements of RNA elements involved in RNA replication', *Nucleic Acids Research*, 40: 6908–21.

Vandamme, A. -M. 2009. Chapter 1: Basic concepts of molecular evolution, in P., Lemey, M., Salemi, A.-M., Vandamme (eds.) *The Phylogenetic Handbook*, 2nd edn, p. 3–29. Cambridge: Cambridge University Press.

Webster, D. P., Klenerman, P., and Dusheiko, G. M. (2015) 'Hepatitis C', *Lancet*, 385: 1124–35.

Wei, L. and Lok, A. S. (2014) 'Impact of new hepatitis C treatments in different regions of the world', *Gastroenterology*, 146: 1145–50.

Weiser, B. M. and Tellinghuisen, T. L. (2012) 'Structural biology of the hepatitis C virus proteins', *Drug Discovery Today: Technologies*, 9: e175–226.

Yau, A. H. and Yoshida, E. M. (2014) 'Hepatitis C drugs: the end of the pegylated interferon era and the emergence of all-oral interferon-free antiviral regimens: a concise review', *Canadian Journal of Gastroenterology and Hepatology*, 28: 445–51

You, S., et al. (2004) 'A cis-acting replication element in the sequence encoding the NS5B RNA-dependent RNA polymerase is required for hepatitis C virus RNA replication', *Journal of Virology*, 78: 1352–66.